MEASURING IMPACT OF STABILIZATION INITIATIVES

TASK 1: DESK REVIEW OF STABILIZATION RESOURCES AND REFERENCES

JULY 11, 2012

This publication was produced for review by the United States Agency for International Development. It was prepared by Management Systems International.

MEASURING IMPACT OF STABILIZATION INITIATIVES

TASK 1: DESK REVIEW OF STABILIZATION RESOURCES AND REFERENCES

MANAGEMENT
SYSTEMS
INTERNATIONAL
A SUBSIDIARY OF COFFEY INTERNATIONAL, LTD

coffey
international
development

600 Water Street, SW, Washington, DC 20024, USA
Tel: +1.202.484.7170 | Fax: +1. 202.488.0754
www.msiworldwide.com

Contract No. AID-OAA-I-10-00002, Task Order No. AID-306-TO-12-00004

Measuring Impact of Stabilization Initiatives (MISTI) Project

DISCLAIMER
The author’s views expressed in this publication do not necessarily reflect the views of the United States Agency for International Development or the United States Government.

CONTENTS

INTRODUCTION AND OVERVIEW

Stabilization initiatives aim to make a country or territory less likely to descend into, or return to a state of conflict or instability, while creating the conditions for long-term sustainable development.[1] International actors have struggled to measure the impact of their stabilization initiatives in conflict-affected environments such as Yemen, Afghanistan, Pakistan, Kenya, or the Democratic Republic of the Congo. Stabilization programming generally takes place in environments characterized by high levels of violent and/or non-violent forms of conflict. Conflict environments are complex. This complexity creates conditions that are dynamic, unpredictable, and subject to problems whose multiple causes and effects may compound one another in vicious cycles of violence and destabilization. Effective solutions to such problems require multi-pronged interventions implemented through an iterative approach for evaluating and learning from past actions to inform future actions.

Conflict-affected environments hinder efforts to evaluate stabilization programming because conditions are unpredictable and rapidly evolving. The complexity of the environment creates a paucity of reliable data because changes on the ground outpace the frequency of data collection, and/or the data needed for accurate assessment change too quickly. Complexity may be further heightened when different actors intervene to change the environment, and the effects of these different interventions interact with each other in ways that are difficult to observe and predict. Complex environments are thus replete with observed and unobserved factors that influence intervention outcomes.

Fortunately, despite such challenges, many donors continue to implement different methods of evaluating stabilization efforts and documenting valuable lessons learned. This report provides an overview of these evaluation efforts and the lessons they offer for programming stabilization interventions, while offering design principles that should be incorporated into future stabilization program evaluations. The first section of this report provides an overview of characteristics of a complex environment. The second section provides background on stabilization programming, its relationship to counterinsurgency, and its challenges. The third section highlights the specific challenges to evaluating stabilization programs followed by a fourth section explaining some of the methods and frameworks that have been used in the past or could be used in the future. The fifth section reviews existing studies and their findings. The report concludes with a proposed way forward for monitoring and evaluating stabilization programming.

CHARACTERISTICS OF COMPLEX ENVIRONMENTS

In recent years considerable attention has been paid to the particular character of the work of international development agencies in environments affected by ongoing violent conflict, or suffering instability in the aftermath of violent conflict. Programs that aim to stabilize conflict environments must effectively manage complexity to prevent conflict in specific areas, and/or mitigate existing conflicts. The complexity of conflict environments is characterized by a high level of uncertainty created by actors with diverse interests, resources, grievances and needs, who interact in continuously shifting patterns of alliance an enmity amid the breakdown of law and order, and social rules and norms governing the legitimate use of force. Unexpected outcomes often emerge from complexity through non-linear processes of cause and effect. For stabilization programming the non-linearity of outcomes has two important implications:

[1] Adapted from USAID's Civil Military Cooperation Policy 2008

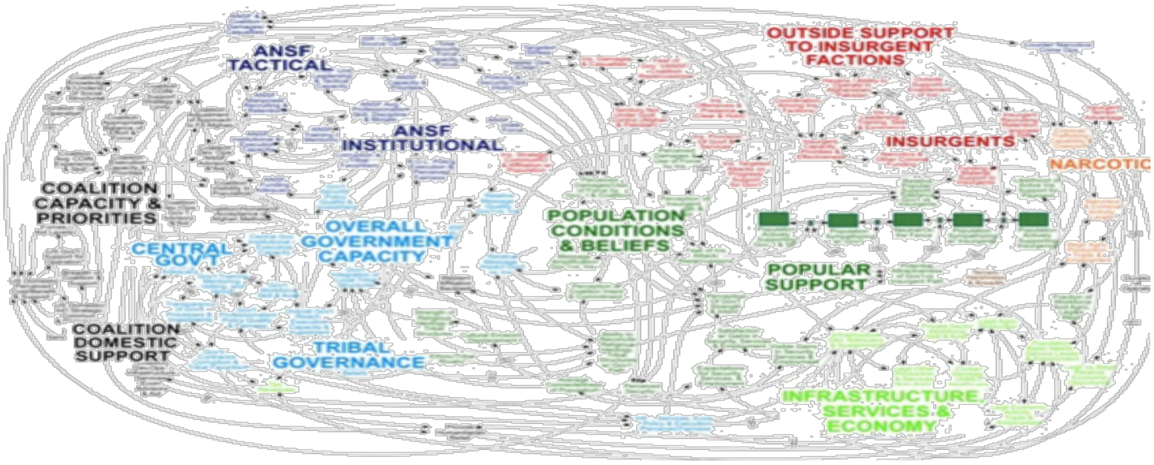
1) The environment does not reliably change in a forward progression from instable to stable. It can become more stable, then reverse and become more unstable, make progress towards stability, and then reverse again. Alternatively, conditions may rapidly shift towards stability or instability as a result of an event such as a peace agreement or other important political shift, the removal or addition of a powerful new actor or actors, a natural disaster, or other sudden change in the physical and social environment that dramatically alters the pattern of social interactions. Eventually a situation may stabilize as a result of focused intervention, but frequent reversals and random events may create difficulty for assessing progress towards stabilization and its causes.

2) Non-linearity makes it difficult to calculate the probability that an intervention will achieve a specific result, and the probability that the result of an intervention will contribute to longer-term objectives. Where there is strong certainty that a specific intervention will achieve a specific result (rarely the case in complex environments), that outcome will cause a set of less predictable effects on other factors in the environment. In a changed environment the same intervention will not achieve the same result a second time, and short-term effects may be much different from long-term effects. To accommodate such complexity, institutions working in complex environments must constantly assess the results of their work and adjust to the changing environment. To determine an appropriate course of action, complex environments require thorough and continual analysis of a range of influencing factors, institutions, and actors.

One of the most important design elements of an evaluation system is an effective methodology for analyzing the environment to inform theories of change and assumptions about program impact. Developing and implementing such a methodology is particularly challenging in conflict environments. Though some organizations have tried to develop detailed influence models to depict the total set of factors driving conflict in Afghanistan, these are often unsupported by empirical data. As a consequence, many analysts have concluded that it is not advisable to attempt to comprehensively diagram an entire conflict system. In fact, such exercises can lead to “analysis paralysis” illustrated by Figure 1 from the US Military Joint Staff.

Figure 1 illustrates how Afghanistan exemplifies an environment where dozens of factors interact with each other in complex and shifting relationships of cause and effect. These dynamics produce a set of “wicked problems”—problems so complex that they have no obvious solution. Attempting to comprehensively diagram a complex environment, such as Afghanistan, can fail to diagnose which factors are more or less important drivers of change.

FIGURE 1: 2009 JOINT STAFF AFGHANISTAN CAMPAIGN DYNAMICS


ANSF TACTICAL
OUTSIDE SUPPORT TO INSURGENT FACTIONS
ANSF INSTITUTIONAL
INSURGENTS
NARCOTIC
COALITION CAPACITY & PRIORITIES
OVERALL GOVERNMENT CAPACITY
POPULATION CONDITIONS & BELIEFS
CENTRAL GOV'T
POPULAR SUPPORT
COALITION DOMESTIC SUPPORT
TRIBAL GOVERNANCE
INFRASTRUCTURE, SERVICES & ECONOMY

Also, sources of instability or drivers of conflict, such as governmental corruption and local grievances, affect specific actors and entities but may not have *systemic* consequences. *Stability or instability is not a characteristic of individual actors or entities. Rather (in)stability describes the state of a system of interactions.* Stability or its opposite is thus a property that emerges from the interactions between the actors and entities that make up a system. In a stable system lines of cause and effect will be relatively easy to identify compared to cause and effect relations in an unstable system, which are relatively non-linear and complex. Non-linear causality is evident in the recognition that conflict creates instability, and instability creates conflict through a complex set of actions and reactions. Understanding the non-linear dynamics of conflict environments is critical because traditional planning tools based on simple chains of cause and effect cannot guide effective intervention. After reaching a working model of the environment the best way to intervene is through a "heuristic" or experiential process of learning through experimentation – conduct an intervention, observe its effects, and learn and adapt accordingly for the next intervention.[2]

Diagnosing Instability Using a Systems Analysis Approach

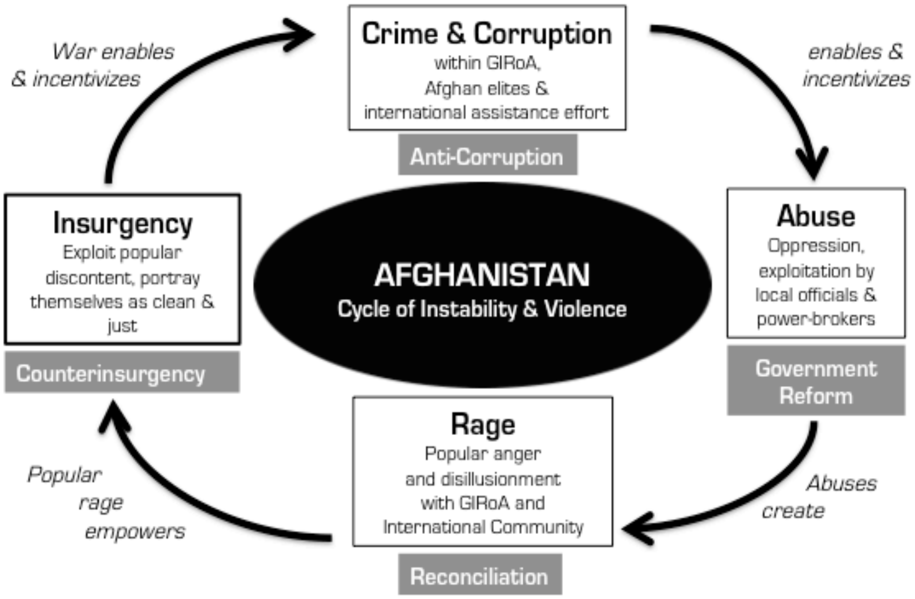
Stability and instability are best understood by taking a dynamic systems approach to analyzing the environment. Donor organizations will lack a clear conception of how to intervene effectively if they do not accurately diagnose the drivers of conflict, and how these drivers affect system dynamics. A dynamic systems approach seeks a coherent understanding of how entities interact to produce a systemic whole that is greater than the sum of its parts. Focusing on types of interaction that create system dynamics enables decision-makers to identify appropriate interventions for changing the pattern of interaction to enhance stability.

The model presented in Figure 2 provides a relatively clear representation of the system dynamics illustrated by the Joint Staff model shown in Figure 1. Instability is created as community conflicts are manipulated by an active and adaptive insurgency, services and security remain poor, and government officials and other powerbrokers accrue wealth and influence by exploiting the population and capturing resources from international assistance. As a result, Afghans' disillusionment and anger with their

[2] USAID Washington is using complexity theory and a complex adaptive systems approach to rethink the traditional program lifecycle in environments threatened by extremism and violence. See Ramalingam, Ben, *USAID's Complexity Journey* in his blog*:* ***Aid On The Edge Of Chaos*** *Exploring Complexity & Evolutionary Sciences In Foreign Aid* http://aidontheedge.info/2011/10/17/usaids-complexity-journey/, accessed on October 26, 2011. For more information on complexity see Sargut, Gokce and Rita Gunther McGrath, *Learning to Live with Complexity: How to Make Sense of The Unpredictable and the Undefinable in Today's Hyperconnected Business World*, Harvard Business Review, September 2011

government and the international community grows, further empowering the insurgency, which creates additional opportunity for crime, corruption, and capturing international assistance.[3] The result is a vicious cycle of increasing instability.

FIGURE 2: A CYCLICAL MODEL ILLUSTRATING INSTABILITY DYNAMICS IN AFGHANISTAN



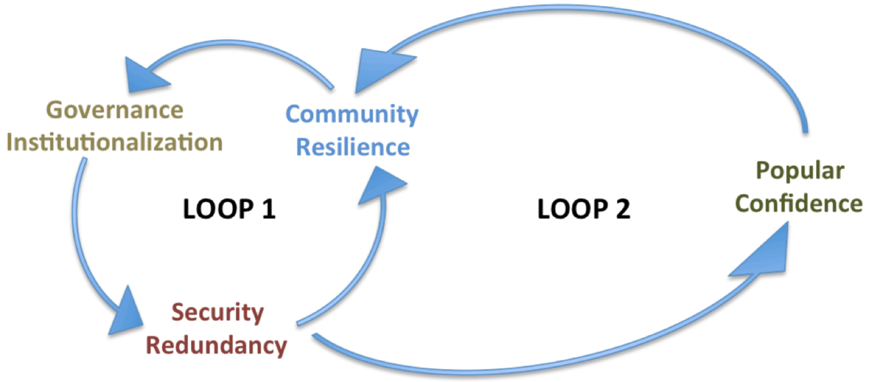
By describing the cyclical nature of system dynamics Figure 2 also demonstrates how interventions must address all parts of the cycle to effectively transform the vicious cycle of instability into a virtuous cycle of stabilization. This type of system diagnosis is useful for showing the main influencing factors behind the system without overburdening the analyst with all of the secondary effects. The diagnosis guides stabilization programs to focus on interventions that are maximally transparent and disassociated with corruption or corrupt officials, that enhance government capacity to deliver services fairly and responsibly, that promote reconciliation between parties to local conflicts, and thereby reduce motives for insurgency.

Where Figure 2 shows the viscous cycle of instability, Figure 3 models the counter tendencies in the system—community resilience, governance institutionalization, security redundancy, and popular confidence—that can be enhanced to produce a virtuous cycle of stabilization. The "double loop" stability model in Figure 3 was developed in 2011 for the 10th Mountain Division in Regional Command South in Afghanistan.[4]

[3] The narrative and corresponding illustration of cycles of instability and violence in Afghanistan is described fully in Dr. David Kilcullen's book *Counterinsurgency*. David Kilcullen, *Counterinsurgency*, (Oxford University Press, 2010), chap. 2

[4] This model was developed by Caerus Associates in partnership with the assessments cell of the 10th Mountain Division in 2011. William Upshur, Jonathan Roginski, and David Kilcullen. "Recognizing Systems in Afghanistan: Lessons Learned and New Approaches to Operational Assessments." Prism. 3. no. 3 (2012, forthcoming)

FIGURE 3: DOUBLE-LOOP STABILITY MODEL



Loop 1 in Figure 3 depicts the key interacting factors that need to be understood locally for designing effective stabilization interventions. Successful interventions should create dynamics of mutual reinforcement between governance institutionalization, community resiliency and security redundancy (Loop 1), which improves popular confidence in non-insurgent institutions (Loop 2), which feeds back to reinforce the dynamics in Loop 1.

The components of the model are measurable through specific indicators: (1) Institutionalized governance is measured by the ability of government institutions to withstand shocks, ability of local councils to make binding decisions, and length of tenure of government officials and community leaders. (2) Community stability and resilience[5] is measured by the ability of local economies and governance institutions to return to normal functioning after a shock created by a violent incident, natural calamity, or other outside factors. (3) Security redundancy is measured by the presence and effectiveness of overlapping government and community security providers, the degradation of the insurgency, and civilian freedom of movement. (4) Popular confidence is measured by population perceptions of security, of local leadership and government, of quality of life, and expectations about future improvements.

Such diagnoses of system dynamics are necessarily (and deliberately) simplified descriptions of intensely complex and nuanced real-world interactions. The method is detailed enough to convey what is important, without being overly detailed to the point of decision paralysis. Not only does a complex systems approach enable more accurate assessment of more and less important areas of focus, but it also helps identify what needs to be changed within the system to stabilize it. This information can then be used to create theories of change that serve as a foundation for the design and evaluation of stabilization programming.

[5] Resilience is referred to in this report as an ability to absorb shocks and rebound quickly.

STABILIZATION PROGRAMMING OVERVIEW

Background

Afghanistan elected its first democratic government in 2004, following nearly 30 years of war and instability. Since coming to office, the Government of the Islamic Republic of Afghanistan (GIRoA) has worked to establish structures of governance at the provincial and district levels, and launch public services that respond to the critical needs of the more than 28 million Afghan citizens.[6]

Continuing violence in many districts exacerbates severe under-development throughout Afghanistan. Insecurity undermines citizen confidence in the legitimacy of the central government and threatens the hard-won gains made to date. The U.S. Government (USG) recognizes the imperative of the nexus of security, governance, and development in stabilizing Afghanistan, and supports GIRoA efforts to establish an effective presence at the provincial and district levels. The approach recognizes that practices and institutions of democracy, especially popular participation in governance, are essential to Afghanistan's long-term development. As a partner with the Afghan people, the USG identifies and addresses local Sources of Instability (SOI)[7] to eliminate the root causes of conflict. The end objective is to establish a stable environment that fosters sustainable social and economic development.[8]

USG's stabilization programs seek to address SOIs by engaging and supporting at-risk populations, extending the reach of GIRoA to unstable areas, providing income generation opportunities, building trust between citizens and their government, and encouraging local populations to take an active role in their development.[9]

Stabilization Programming's Relationship to Counterinsurgency (COIN) Efforts

The USG's "whole of government" approach gives USAID an important role in civilian and military COIN efforts. Military COIN efforts aim to support the establishment of government control over areas where insurgents are active, while reducing popular support for insurgents and the capacity of the insurgency to operate. USAID stabilization programming contributes in the short and medium term to political and social cohesion, community resilience, and better governance—all essential to enable areas "cleared" by kinetic military action to be held securely, denying insurgents the possibility of drawing support from the local populace.[10]

Counterinsurgency generally includes four phases: shape, clear, hold, and build.[11] The "Shape" phase includes efforts aimed at assessing an area to identify SOIs, and then addressing those SOIs in an attempt

[6] US Mission Afghanistan Performance Management Plan, Annex V11 – Assistance Objective 7: Stability Sufficient for Basic Governance and Sustainable Development, April 2011

[7] Sources of Instability are local factors that: 1) Decrease support for GIRoA; 2) Increase support for Anti-Government Elements (AGEs); and 3) Disrupt the normal functioning of society. See the District Stability Framework Student Book Version 3, August 2011

[8] US Mission Afghanistan Performance Management Plan, Annex V11 – Assistance Objective 7: Stability Sufficient for Basic Governance and Sustainable Development, April 2011

[9] Ibid.

[10] US Mission Afghanistan Performance Management Plan, Annex V11 – Assistance Objective 7: Stability Sufficient for Basic Governance and Sustainable Development, April 2011

[11] Petraeus D.H. and Amos, J.F., Counterinsurgency, Headquarters of the Army, Field Manual FM 3-24 MCWP 3-33.5,

to reduce susceptibility to insurgent influence. For stabilization programs “Shaping” activities usually involve engaging local community leaders and key local influencers in order to identify grievances that insurgents can use to gain popular support. Once identified, these grievances can be addressed to prevent insurgency.

The “Clear” phase is the first phase of COIN engaged in areas of Afghanistan where groups are already conducting violent anti-government activities. This phase involves military operations to remove insurgents from the area. Stabilization efforts in this phase include quick impact activities to meet recovery needs in priority communities and assist local government entities to establish or strengthen their presence.[12]

The “Hold” phase starts when insurgents have been cleared and the population has been secured. This phase involves rapid response to any new security threats, and quick-impact activities to win the support of the local population. Stabilization activities in this phase are focused on addressing SOIs by enhancing community resilience, improving governance, and increasing popular confidence in GIRoA through integrated community development projects.[13]

The “Build” phase involves medium- to long-term efforts to ensure durable stability in an area. Stabilization efforts during this phase focus on rebuilding key infrastructure, improving government capacity, and expanding the delivery of government services. Such activities support a transition from stabilization to long-term development.

Transition is now at the center of USG strategy with the handover of security responsibilities from the International Military Assistance Force (ISAF) to the Afghan National Security Forces (ANSF) scheduled for completion by the end of 2014. With the transition in security arrangements the main objective of stabilization programming is to create conditions for Afghan-led sustainable development to take place.[14]

USAID Mission Afghanistan Stabilization Programming

In February 2010, USAID formed the Stabilization Unit to gather all USG stabilization programs and planning capacity under one office, and assigned staff assigned to coordinate stabilization with national development programming managed through other technical offices. The unit is responsible for addressing and responding to USG stabilization objectives and priorities, managing stabilization programs, representing the USAID in civilian-military coordination with the U.S. military and the International Security Assistance Force (ISAF), and socializing the principles of stability programming with key stakeholders in GIRoA and the USG.[15]

The Stabilization Unit has designed four stabilization programs to operate in geographic areas that fall into the different phases of the stability continuum – clear (red), hold (orange), build (yellow) and

[12] US Mission Afghanistan Performance Management Plan, Annex V11 – Assistance Objective 7: Stability Sufficient for Basic Governance and Sustainable Development, April 2011

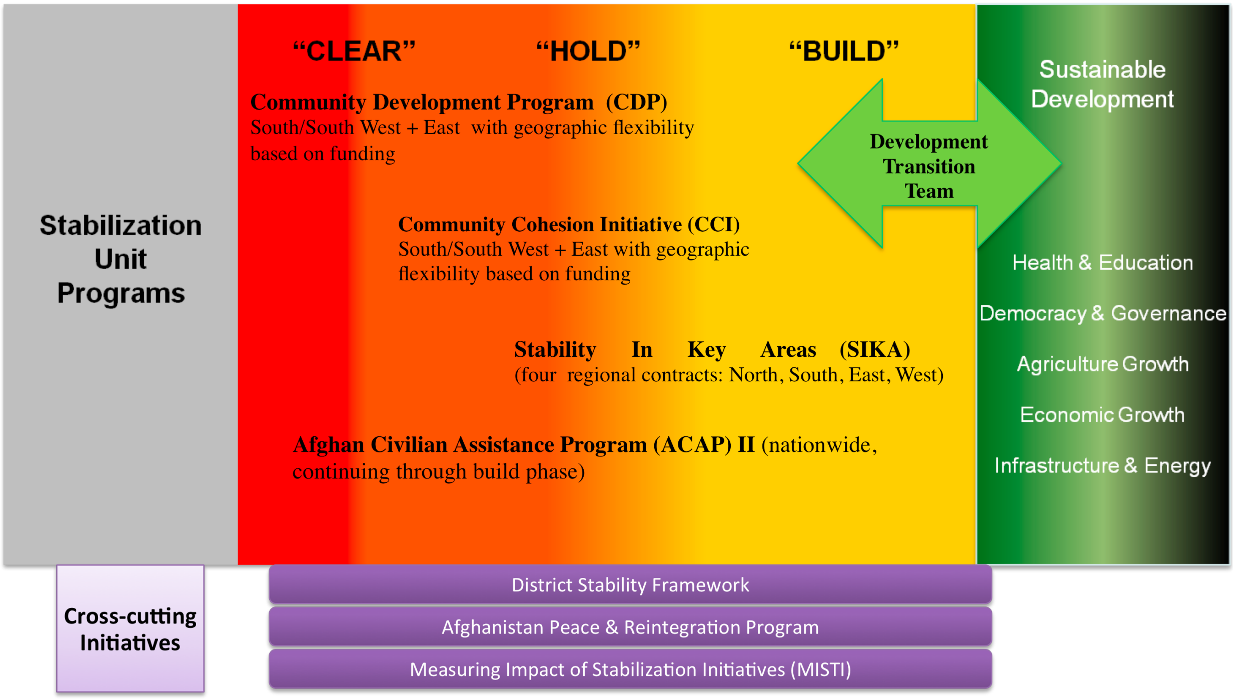
[13] Ibid.

[14] In the last few years, USAID has distinguished among assistance objectives and approaches that inform traditional development programming and those that guide the Agency’s stabilization or COIN initiatives. See USAID Policy, "The Development Response to Violent Extremism and Insurgency." Last modified September, 2010. Accessed May 15, 2012 http://www.usaid.gov/our_work/policy_planning_and_learning/documents/VEI_Policy_Final.pdf

[15] US Mission Afghanistan Performance Management Plan, Annex V11 – Assistance Objective 7: Stability Sufficient for Basic Governance and Sustainable Development, April 2011

transition (green) shown in Figure 4.[16] The four programs – Community Development Program (CDP), Community Cohesion Initiatives (CCI), Stabilization in Key Areas (SIKA), and the Afghan Civilian Assistance Program II (ACAP II) – are arrayed on the stabilization continuum in Figure 4 according to their geographic focus.

FIGURE 4: USAID/AFGHANISTAN'S STABILIZATION PROGRAMS AND THE COIN CONTINUUM


Stabilization Unit Programs
"CLEAR"
"HOLD"
"BUILD"
Community Development Program (CDP)
South/South West + East with geographic flexibility based on funding
Community Cohesion Initiative (CCI)
South/South West + East with geographic flexibility based on funding
Stability In Key Areas (SIKA)
(four regional contracts: North, South, East, West)
Afghan Civilian Assistance Program (ACAP) II (nationwide, continuing through build phase)
Development Transition Team
Sustainable Development
Health & Education
Democracy & Governance
Agriculture Growth
Economic Growth
Infrastructure & Energy
Cross-cutting Initiatives
District Stability Framework
Afghanistan Peace & Reintegration Program
Measuring Impact of Stabilization Initiatives (MISTI)

The Community Development Program (CDP) responds to the exigencies of the immediate post-clear context through activities that are prioritized by the local community and planned and implemented in coordination with GIRoA officials to the greatest extent possible, depending on the situation in the local area. CDP implements focus highly visible infrastructure repairs, temporary employment for large numbers of local Afghans, and on-the-job skills and community maintenance training to improve livelihoods in the wake of military operations.

The Community Cohesion Initiative (CCI) works in more permissive environments, most often falling into the hold phase of the stability continuum. CCI seeks to increase community resilience in areas vulnerable to insurgent exploitation by (1) strengthening ties between local actors, customary governance structures, and GIRoA, and (2) increasing cohesion among and between communities. Using a process-oriented, community-driven approach, CCI utilizes clusters of small grants including both "soft" and "hard" activities. Soft activities include community leadership shuras, District Governor outreach visits, and other relationship building activities. Hard activities include small-scale infrastructure repairs that bolster local level productivity and capacity.

[16] Afghanistan Strategy: 2012-2015, Office of Transition Initiatives, June 2012

The Stabilization in Key Areas (SIKA) program partners with the Ministry of Rural Rehabilitation and Development at the national, provincial and district levels to enhance governmental capacity and coordination in an effort to link informal community and sub-district governance institutions with formal district, provincial and national government institutions. SIKA focuses in the build/transition phase of the stability continuum to consolidate stability in key areas such that longer-term sustainable development programming can take place.[17]

The Afghan Civilian Assistance Program II (ACAP II) supports the effort of innocent Afghan families to recover after suffering from violent incidents where international military are involved, thereby reducing instability arising from war-affected families. ACAP II operates across all phases of the stabilization continuum. Beneficiaries receive direct, immediate in-kind assistance such as medical treatment, household goods, agricultural equipment, and livestock, as well as longer-term vocational training or livelihood support in eligible cases. The program may also refer families to local NGOs who can provide necessary services.

Stabilization Programming Challenges

Short Time Horizons

Stabilization programs are often expected to demonstrate impacts over short periods of time, sometimes as little as three to six months. As a result, programs must prioritize what they can do quickly, rather than implementing longer duration activities that may be able to achieve more significant results. As described in the next section, there remains a tension between the requirement to demonstrate quick impacts, and the desire for analytically rigorous measurement approaches that require a time-intensive design process in advance of the initiative that is being evaluated. Relatively expedient measurement techniques are often leveraged, sometimes at high cost, which provide only vague indications of program impacts because of a less-than-rigorous evaluation design.

Challenge of Developing a Common Understanding of the Environment

Another challenge of stabilization programming is ensuring an accurate diagnosis of the system and understanding of the environment. If speed is the priority, a project's end state risks being reduced to the completion of the project itself, rather than a demonstrable effect. This reductionism is a symptom of incapacity to define a shared diagnosis of the system such that there is a clear understanding of how an intervention that changes one of the system's components is most likely to affect the other components to change the system as a whole. Without a shared understanding of the system no realizable end state can be identified that would result from stabilization program interventions. A good example of this problem is illustrated by the Commanders Emergency Response Program (CERP) in Afghanistan, where commanders have an incentive to spend quickly and massively, often without pre- or post-assessments of impacts. Efforts to synch CERP with GIRoA priorities has been case study of mismatched understandings of the environment and unrealistic expectations on all sides. Further, the temptation with CERP is often to "build things", not build people or Afghan capacity – a far more strategic and enduring stabilization objective.

Organizations – civilian, military, coalition, NGO – must nevertheless develop a common understanding of their environment as a dynamic system, within which they can identify the critical dependencies driving instability and resiliency. Yet, in a recent "meta-evaluation" of previous USAID evaluations, only 26% percent of evaluators "examined the conceptual framework of the logic model underlying the

[17] Ibid.

evaluated development intervention in order to clarify the causal relationship between inputs, outputs, activities, outcomes, and impacts." [18]

Challenge of Learning and Adapting to Changing Environments

Working in a complex dynamic system requires that stabilization programming constantly learn and adapt to the environment. This is especially true for any U.S.-led initiatives as U.S. forces and civilians have traditionally been deployed to relatively unstable areas. Program objectives and operational approaches must be able to adapt as the environment changes. Over time, implementers and evaluators should cultivate a progressive understanding of the kinds of intervention that catalyze systemic changes. Implementers must work closely with M&E specialists to ensure that programs are adaptable, while still maintaining an explicit theory of change, grounded in a broader theory of the environment.[19]

CHALLENGES OF MEASURING STABILIZATION INITIATIVES

Measuring progress in any complex environment — in security, development, governance, and overall stabilization — has proven extraordinarily difficult. Identifying the key challenges to measuring progress is the first step for developing appropriate and effective systems for evaluating stabilization programs. The complex environments where stabilization programs are implemented exhibit a high degree of uncertainty with rapidly changing dynamics, making specific outcomes virtually impossible to predict except in the immediate term, complicating measurement regimes. These environments further exacerbate operational vulnerabilities such as poor situational awareness, a lack of reliable data, short timelines for project design and execution, ambitious objectives, unexpected changes in priorities, limited direct observation and episodic monitoring of projects, potential conflicts of interest created by implementing partner self-reporting, and limited partner capacity for assessment and evaluation. These obstacles create a tension between analytic rigor and operational constraints because analytic rigor may be too expensive or close to impossible to achieve, or the pressures to respond to a changing environment does not allow the time to set up the structures needed in advance (such as a baseline) for an analytically rigorous evaluation process.

Though complexity science is beginning to influence development practice, it remains the exception, not the norm. At present, many M&E approaches do not embrace the principles of complexity. Instead, they continue to rely on snapshots in time informed by single-source or single-methodology approaches, such as key informant interviews or basic quantitative analyses of polling data using summary statistics rather than explanatory or predictive modeling.

In addition, the non-linear dynamics of a complex environment require a systems-thinking approach to monitoring and evaluation. Systems-thinking enables an adaptive understanding of the environment. To this end, the author and USAID advisor, Michael Quinn Patton, has written extensively on the benefit of monitoring and evaluation (M&E) systems designed around the insights of complexity science, encompassing an interdisciplinary approach that takes into account the overall behavior and decisions of all actors in the area under exploration. His 'developmental evaluation' framework prioritizes an understanding of local context, responsiveness, and adaptation, challenging M&E specialists to

[18] Office of the Director of U.S. Foreign Assistance, "A Meta-Evaluation of Foreign Assistance Evaluations." Last modified June 1, 2011. Accessed May 15, 2012.http://pdf.usaid.gov/pdf_docs/PCAAC273.pdf

[19] For more information on lessons of counterinsurgency programming as part of stabilization programming, see USAID/OTI Lessons Learned in Counterinsurgency http://transition.usaid.gov/our_work/cross-cutting_programs/transition_initiatives/lessons_coin.html

conceptualize evaluation as a catalyst for systemic change rather than a tool for improving specific program efforts.[20]

However, the systems-thinking approach to M&E is challenging to implement in reality. In Afghanistan, one of the challenges is having up-to-date and accurate analysis of unfolding events and the impacts of interventions at the local level. Many of the evaluation approaches in Afghanistan and other complex environments were designed to respond to less fluid environments and analysis an environment at a national or regional level, rather than a local level. As a result, they can be unwieldy and difficult to get quick results.

In Afghanistan, organizations' frameworks for learning should be continuously tested and updated, yet this has been a slow and disjointed process. While donors and ISAF have tried to monitor the outcomes and progress of their development assistance and security efforts, civilian and military planning and operations have not proven well-adapted for tracking the highly localized and evolving environment of counterinsurgency (COIN), stabilization, and transition. This challenge is compounded and exacerbated by rapid turnover of staff and regular rotation of military units. It is often said that Afghanistan has been 10 one-year wars – not one ten-year war. Moreover, many stabilization and COIN programs only analyze their own activities. Ideally a detailed understanding of an insurgency's system of competitive control and its impact on communities would inform all stabilization and COIN programs. However many actors do not have a nuanced understanding of insurgency dynamics for lack of information or analysis. Communication barriers between military and civilian classified systems and their implementing partners inhibit the transfer of information to implementing partners and from implementing partners to the broader policymaking apparatus. Glaring cases of disconnects between U.S. civilian agencies on the ground and their military counterparts Illustrate how the cross-cultural challenges that require the most attention are sometimes not relationships with foreign nationals.

Developing Coherent Theories of Change

Coherent theories provide the basis for the design, implementation and evaluation of USAID programs. Theories of change help to identify or predict causal linkages between inputs and outcomes. Effective programming requires that theories of change are identified, made explicit during program design and implementation, and regularly tested through evaluation. Additionally, programmatic theories of change are much more useful if they are informed by an overall theory of the environment, giving program staff a sense for the territorial or system logic that helps them understand "how things work" locally.

It can be time consuming and analytically challenging to define theories of change in the context of a dynamic systems-oriented understanding of the environment. Pressures to spend large amounts of money, expedite program roll-out, and demonstrate impacts along with conflicting guidance can threaten program coherence. In April 2012, a portfolio review of USAID/Afghanistan's Stabilization Unit revealed that several programs being implemented did not have clear guiding theories of change. Project managers, implementers, and other stakeholders spoke about the general importance of having a defined theory of change for their projects and many have had them implicit in their program design, but they are frequently not articulated in a way that is testable and measurable. Programs with explicit theories of change sometimes lack a broader theory of the environment, preventing decision makers, managers, and implementers from developing a common picture of patterns of instability and resiliency in their program areas.

[20] Patton, Robert Quinn. Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use. New York: The Guilford Press, 2011.

Attribution Dilemma

In a complex system, multiple actors, institutions and other factors influence the environment and behavior. The sheer variety and density of international civilian and military organizations operating across the South and the East of Afghanistan, each of which is pursuing its own objectives and conducting its own efforts, makes measuring the effects of one program and attributing causality to a specific intervention particularly difficult to realize. In these cases, plausible attribution may be more a more realistic objective than verifiable causality. Any evaluation demonstrating causal attribution has to clearly identify the other factors might have contributed to or hindered the same impact.

Aggregation Dilemma

Any regional or national measurement analysis must also address the challenge of aggregating local information into meaningful macro patterns. The process of aggregating village, district or provincial information into a broader trend analysis can be easily manipulated to disguise or delete outlier data points.[21] Aggregation generally dilutes or obscures local context, reducing complex data to a simple chart or dashboard and framing the underlying patchwork of micro-environments out of the analysis. Unfortunately such manipulation may be an intentional part of an effort to brief positive outcomes. Typically, when military units rotate, the incoming unit assesses the majority of its area of responsibility (AOR) as insecure (or red) so that when it departs, it can demonstrate progress toward stability with a larger part of its AOR categorized as more stable (yellow or green color shading on the COIN continuum). For example, an April 2012 trip to RC-East revealed that the outgoing Division's final briefing on stability trends in the region deleted critical data from highly insecure districts, allowing the division to show a continuously improving trend line. As a result of this behavior, it is common for important information about highly unstable districts to be ignored in the average stability score for a province, or for a highly unstable province to be averaged out of a regional assessment.

The aggregation dilemma is particularly acute in Afghanistan, where the environment is best understood as a highly localized mosaic of micro-environments. Programs that are designed to be adaptive to local context are particularly challenged by the political imperative to demonstrate overarching stability impacts. For example, USAID's Office of Transition Initiatives (OTI) in Kabul and its implementing partners for the Afghan Stabilization Initiative (ASI) were consistently pressured to provide a summary analysis of overarching stability impacts across all levels (from village up to the national level). In an April 2012 meeting in Kabul, OTI's M&E team revealed that due to a lack of capacity they had to largely rely on evaluators' synthesized analysis, making their summary analysis susceptible to the aggregation dilemma. Multilevel methods of analysis are required to show the nuances within an environment in order to resolve the aggregation dilemma. It is also worth stressing that the Afghan voice might best be translated not in documents, but by inviting Afghans to higher-level, national sessions when assessments are under review.

Evaluations Require Assessing Changes in Perceptions and Behaviors

One of the significant challenges with evaluating stabilization programs is that the desired impact is usually a change in perceptions and behaviors. Evaluating these changes is challenging. Perceptions generally require focus groups, key informant interviews and/or mini or large-scale individual surveys. Although very useful, these tools can be time consuming, and there are more types of biases that can appear in the data compared to other types of programs whose impact can be more readily measured by

[21] Connable, Ben. RAND Corporation, "Improving Counterinsurgency Campaign Assessment: The Importance of Transparency in the Fog of War." Last modified April, 2012. Accessed May 15, 2012. http://www.rand.org/content/dam/rand/pubs/research_briefs/2012/RAND_RB9645.pdf

counting increases in particular outputs that result from an intervention. Additionally, it can be challenging to determine the correct metrics for assessing changes in perceptions and behaviors.

Lack of Baseline Data

Baseline data against which to track trends and the impact of projects often does not exist in Afghanistan. This is frequently due to the pressures to begin programming right away rather than waiting for a baseline to be undertaken. Where such data does exist, it is sometimes of questionable quality and has little utility to the researcher. Most historic data is ad hoc or infrequent. While donors and ISAF have tried to monitor the outcomes and progress of their development assistance and security efforts, civilian and military planning and operations have not proven well-adapted for tracking the highly localized and evolving environment of counterinsurgency, stabilization, and transition.

METHODS, FRAMEWORKS, AND DATA

The challenges and costs involved in measuring stabilization impacts in Afghanistan and other high-threat complex environments deter organizations from investing in effective impact evaluation and measurement systems. Traditional approaches consistently fall short in accurately determining impact in complex environments. For example, a 2011 report by the International Initiative for Impact Evaluation (3ie) determined that among 165 U.S. stabilization interventions, only 44 projects had conducted evaluations, only one of which met a high standard of rigor.[22] Some organizations, such as the World Bank and the United Nations, have been more aggressive, conducting randomized impact evaluations in Afghanistan and elsewhere, but even these approaches can lack the rigor needed to reliably measure impact and inform future resource allocation and planning.

Methodological Approaches

Impact evaluations vary along a spectrum of analytic rigor, ranging from anecdotal approaches to those that are experimental. The main categories in order of rigor include:

- **Anecdotal:** Where one observes behavior as it appears or becomes known, records it, and uses various quantitative and qualitative methods to analyze the gathered data and derive an understanding of the observed activity. Rigor is compromised because there is not a strong methodology determining which behaviors are selected for observation and analysis, and how the latter is accomplished. As a result, one cannot control for selection bias or establish causal effects. Anecdotal impressions are unreliable, often idiosyncratic, and potentially misleading, making them difficult to defend.

- **Systematic:** Makes use of purposive or random sampling or observations designed to be representative of a particular population or set of phenomena but does not necessarily incorporate control and treatment designs to establish counterfactual cases. Data collection is systematic and correlations between the data are established. Systematic analysis may enable the plausible attribution of observed effects to specific interventions and allow testing a program’s theory of change for validity. Selection bias and unobserved heterogeneity bias arising from the multiple

[22] Brown, Annette. International Initiative for Impact Evaluation, "Evaluating Stabilization Interventions" Last modified July 20, 2011. Accessed May 15, 2012. http://www.3ieimpact.org/userfiles/doc/Evaluating Stabilization Interventions July 2011 Delhi presentation.pdf

influences of a complex environment remain factors that reduce confidence in the attribution of specific effects.

- **Quasi Experimental:** Is used when researchers cannot control the allocation of the treatment being studied. In other words, the researcher cannot select the group receiving the treatment and the control group that does not receive the treatment. This affects the experiment's internal validity. To a degree one can control for selection bias, though typically only for observed variables. Advantages over anecdotal and systemic research include its precision and the ability to apply results to other settings in addition to establishing a counterfactual case. However, it is still less rigorous than experimental design because there is not complete control for selection bias.

- **Experimental/Randomized Control Trails (RCTs):** RCTs allow researchers to randomly select treatment and control groups, eliminating selection bias and improving internal validity. RCTs also provide "clean" estimates of causal effects, improve cost effectiveness, and provide defensible and generalizable findings for evidenced based policy-making. RCTs may not however be applicable to stabilization programs where the onus is on the flexible and strategic allocation of resources and quick reaction time to changing events and priorities created by the complex environment. Programs may thus perceive rigid treatment and control designs as allowing the evaluation "tail" to "wag the dog" of the program.

Frameworks

In Afghanistan, two frameworks have been used for analyzing and evaluating stabilization programming and its impact on the environment: The District Stability Framework and the Region South Stability Approach.

- **District Stability Framework (DSF)**: Originally pioneered by USAID's Office of Civilian-Military Cooperation, the DSF is primarily an analysis and planning framework, but it was also designed to support monitoring and evaluation. Used as a common tool across USG civilian and military actors primarily in RC-East, the DSF is used to analyze the environment in order to identify sources of stability and instability. Stabilization programming is then designed by both civilian and military actors to address the sources of instability and bolster resiliencies. The framework includes a monitoring and evaluation matrix for progress towards stabilization and how it will be measured.[23]

    - **Strengths:** DSF provides a consistent framework and language that both civilian and military officials can use so that efforts are jointly coordinated and evaluated. Particularly, it gets actors to focus on sources of instability and stability as a start of analysis. It is particularly effective with USAID implementing partners that can use it as a framework from which to base their monitoring and evaluation efforts. It is relatively clear, provides a starting point for people to do the accurate analysis needed on which to found their theories of change, and local national staff have found it particularly helpful in helping them to analyze a district's political, economic and social situation.
    - **Weaknesses:** The DSF does not incorporate a broader systems-based analysis of the environment so the focus on sources of instability and stability can be limiting. It is also quite complicated with its series of matrixes that can be very time consuming to fill out. As a result, the DSF is not used uniformly across Afghanistan. Moreover, the monitoring and evaluation matrix tends to be its weakest part and is not used as frequently as the

[23] For more information on the District Stability Framework, see http://usacac.army.mil/cac2/call/docs/11-16/app_b.asp

other parts of the framework. Even when the monitoring and evaluation matrix is utilized, the quality, timeliness, and availability of data to support those efforts varies and it does not push its users to analyze progress made against those indicators. Lastly, assessments conducted by brigades and battalions can be subsequently lost as units are replaced in theater.

- **Region South Stability Approach (RSSA):** The RSSA is unified framework that is used to analyze where that particular district is on a stability continuum from instable to stable. It relies on a matrix that demonstrates the characteristics of a district at each phase of the shape, clear, hold, and build continuum. These characteristics translate into the objectives against which civilian and military actors focus their efforts in that district. For example, if a district is in the "clear" phase and one of the characteristics of the "hold" phase is that a legitimate local decision-making body is constituted, then stabilization efforts will focus on the objective of supporting the establishment of that local decision-making body.
  - **Strengths:** It is widely used across RC-South as a common tool for civilian and military personnel and includes quarterly information collection that helps to provide regular reliable information. It is relatively simple to use.
  - **Weaknesses:** The RSSA uses a continuum approach that does not fit within systems-oriented approaches. It also segregates by sectors (Agriculture, economics, etc.), which makes it more challenging to develop an integrated approach. Some of the indicators are subjective and vulnerable to corruption. Frequent personnel turnover makes it harder to ensure consistent application.

- **Interagency Conflict Assessment Frameworks:** USAID's Office of Conflict Management and Mitigation pioneered the Conflict Assessment Framework (CAF 1.0 and 2.0) which was later developed into the Interagency Conflict Assessment Framework (ICAF) with the State Department's Office for Coordination, Reconstruction and Stabilization (now the Bureau for Conflict and Stability Operations).[24]
  - **Strengths:** These tools have catalyzed exploration of conflict-sensitive theories of change[25], which have begun to inform the design, implementation and evaluation of USAID's conflict mitigation programming. The ICAF provides a common platform for interagency analysis to serve as a foundation for monitoring and evaluation efforts.
  - **Weaknesses:** Both the CAF and the ICAF represent snapshots in time of the key drivers and mitigating factors of conflict. Conflict is a dynamic, evolutionary process that cannot be captured or described in a single analytical effort. Also, these tools focus at the macro, national level, which makes it challenging to account for the diversity of conflicts across Afghanistan, each of which is highly localized.

One of the strengths of these frameworks as a whole is that their presence and utilization has helped to catalyze exploration of conflict-sensitive theories of change. However, consistency is lacking in the definition and application of these concepts across frameworks. These frameworks are not interoperable and do not directly inform one another. Both RSSA and DSF do attempt to mitigate the aggregation dilemma by relying on the expertise and perspective of USG and NATO partner-country counterparts (at

[24] For more information on the Conflict Assessment Framework see https://dec.usaid.gov and for the Interagency Conflict Assessment Framework see http://www.state.gov/documents/organization/187786.pdf

[25] For more information on conflict sensitive theories of change and corresponding indicators see USAID Theories of Change and Indicator Development in Conflict Management and Mitigation. June 2010

the DSTs, PRTs, and RCs/RPs) to build the assessments and synthesize relevant patterns while incorporating their contextualized understandings. However, the experiences and insights of foreigners is vulnerable to observation and selectivity bias, especially when subject matter experts are subjectively reporting progress against objectives for which they are accountable. They have tried to mitigate this in some places by having local staff input from USAID implementing partners in the DSF frameworks.

Dynamic Treatment Regimes

Typically, stabilization assistance is not a one-time event but rather a sequence of multiple interactions and interventions between donor and recipient. Much of USAID's current stabilization efforts are built on a sequence of assistance in the Clear/Hold/Build/Transition framework. One way to ensure that this sequence logically progresses over time is to use the Dynamic Treatment Regime (DTR) to determine treatment choices based on the effectiveness of prior treatments.

DTR are a set of rules for choosing the administering of a "treatment" to individual "patients" (i.e. villages). The medical language is intentional, as DTR have been embraced by the medical community as the "platinum" standard for determining how to administer treatment choices for particular patients based on their individual characteristics and history. DTRs emerged from the realization that medical interventions are not typically one-time events but rather involve a sequence of treatments.

To analyze the interaction between aid programs over time, the DTR would use a multistage approach where randomization is employed each time aid is administered. Rather than "blind" randomization, however, the implementers would shape the randomization process by dictating the proportion of villages (but not the individual villages themselves) to receive a certain type of treatment.

A typical DTR would follow these steps:

(1) Conduct an assessment of the sample population (i.e. the villages to be considered)
(2) Assign those villages to one of two types of treatment (the treatments provided are tailored to the objectives of the program., i.e., small grants, job programs, etc). Re-evaluate those villages at some point after receiving the first round of treatment (i.e., at 6 months)
(3) Administer a second round of treatments (again, there's flexibility here, and could be more of the first type of treatments or new ones), again randomly
(4) Re-evaluate outcomes.

The DTR framework has several notable advantages over a one-shot experimental design. First, unlike other designs, it explicitly models the interaction between different types of assistance over time. As a result, one is able to determine which combinations of assistance are most likely to achieve the desired outcome. This moves us much closer to "tailored assistance" than possible with other designs since we also have information on which sequence of assistance (not just which type) is most effective given a village's background characteristics. DTRs thus provide information that is of direct use in planning tools for future rounds of assistance and in other areas where similar scope conditions apply.

Second, it forces one to have both a clear idea of the outcome of interest (i.e. what is the aid trying to accomplish) and the measurement strategy, which must be uniformly applied across the villages. This helps ensure analytical rigor and facilitates comparison to other areas in (and beyond) Afghanistan. Finally, it draws on the power of randomization at multiple stages, shielding the design from bias due to unmeasured (or omitted) variables and controlling for confounding factors due to a village's prior history, geographical location, or socioeconomic profile.

Example: Imagine a stabilization program administered in 100 villages that an assessment process has identified as being in need of assistance. The outcomes of interest are community resilience and improved attitude toward the district leadership. Imagine that there are four treatments we'd like to examine:

- Treatment A: a small grant
- Treatment B: a jobs program
- Treatment C: a larger grant
- Treatment D: doing nothing

In Round 1, 50 villages would be assigned to Treatment A and 50 to Treatment B. A survey would then be conducted at the six month mark to measure community resilience and attitude toward the district leadership. In Round 2, 50 villages would be assigned to Treatment C and the remaining 50 villages to Treatment D. The evaluation process would then be repeated in another six months after aid in Round 2 had been assigned (a full year now after the aid program began).

In this simple setup, one would be to evaluate four possible aid regimes: AC (small grant + a bigger small grant); AD (small grant + doing nothing); BC (job program + a bigger small grant); and BD (job program + doing nothing), enhancing our understanding of the interaction and dynamics of aid programs over time in conflict settings by providing analysis after each round.[26]

Other Tools

Within these methods, there are different tools that can be used to support evaluations. These include key informant interviews, basic quantitative analysis of polling data, capturing of stories and anecdotes, and focus groups. One example of complexity design that has been used in Afghanistan to test COIN theory is the Defense Advanced Research Project Agency's (DARPA) Computational COIN Program. This program employs a combination of remote observation, deep dive qualitative interviews augmented by household surveys, and multi-level modeling of diverse data sets in support of stability assessments for Afghan provinces.[27] Mixed data collection and analytical methods embedded in a systems analysis of the environment (i.e., the dynamic interdependencies), is a powerful combination that should be designed and executed in parallel with project design and implementation. However, the challenge to date is that many of these efforts are taking place in classified environments, preventing many stabilization actors from contributing to or gaining from the information. This will continue to be a problem until the communication barrier between those working stabilization programming in an unclassified environment, and those working in a classified environment can more easily transcribe information back and forth as appropriate.

Useful new approaches are now being developed that combine quantitative methods and remote observation with on-the-spot qualitative field research by teams of local observers or researchers. Methods include community-led monitoring and evaluation, in depth-interviews, focus groups, crowd-

[26] Jason Lyall, Dynamic Treatment Regime, May 26, 2012

[27] The Defense Advanced Research Project Agency's (DARPA) Computational COIN program leverages cultural intelligence in support of stability assessments for Afghan provinces. The program processes existing data sources in non-traditional ways, enabling the team to test 50-year old standing COIN theories, and then modify, adapt, and optimize them. Using remote data collection, analyses can be performed on areas that have little Coalition presence; "observer effects" are minimized and stability indicators are scalable. Supplementing this computational social science research, an Afghan field research network provides specific qualitative analysis to corroborate other hypothesized relationships amongst remotely collected data, such as district analysis, mapping of powerbrokers, and various manifestations of control – such as frequency and permanence of insurgent courts. Local researchers provide unique insights on issues including local perceptions of security, support for government initiatives and institutions, conflict dynamics and local disputes.

sourced reporting technologies such as participatory mapping, and SMS surveys.[28] Fusing these perspectives allows researchers to develop an integrated and contextualized picture of conditions on the ground.

Data Sources

An important component to support these frameworks or any reliable and effective monitoring and evaluation system is reliable, regular, and relevant data. The list below describes selected data sources reviewed for this paper. The data sets are described in terms of their applicability to the MISTI project, the frequency with which they are updated, and the reliability with which their information can contribute to monitoring and evaluation efforts under the MISTI framework.

DATA SOURCE	APPLICABILITY	FREQUENCY	RELIABILITY
District Stability Framework (DSF)	While DSF is built around Sources of Instability, it is a planning framework that is not well suited to track overall stability trends.	BINNA is a nation-wide household survey conducted quarterly by the ISAF Joint Command (IJC). It is the primary data source for DSF though other country-wide surveys such as ANQAR may also provide inputs to DSF. BINAA is now in its fourth round of data collection in RC-E. Additionally there is no good indication of how widely or robustly DSF is being implemented in RC-S.	The BINAA survey is long, subject to invalid aggregation techniques and inadequate sampling sizes. It is generally a poor data source.

[28] Reproduced from Kilcullen, David and Alexa Courtney, "Big data, small wars, local insights: Designing for development with conflict-affected communities" in What Matters, McKinsey and Company, December 2011 accessed on May 16, 2012 at http://whatmatters.mckinseydigital.com/social_innovation/big-data-small-wars-local-insights-designing-for-development-with-conflict-affected-communities.

DATA SOURCE	APPLICABILITY	FREQUENCY	RELIABILITY
Region South Stability Approach (RSSA)	RSSA utilizes a continuum approach that does not fit within systems-oriented approaches. The RSSA also segregates discreet sectors (E.g., Agriculture, Economy, etc.). It will be of limited use informing a more integrated systems approach.	Given its wide application in RC-S and regular, quarterly collections, RSSA can serve as a reliable source of data.	A substantial portion of RSSA indicators and metrics are subjective and vulnerable to corruption. Frequent personnel turnover in the RC compounds the issue.
NRVA	NRVA tracks general sector-specific metrics and indicators at the regional, provincial, district and village level.	Given its country wide, multi-level focus, NRVA is only completed every few years. For stabilization operations, frequent data collections are required more frequently to effectively monitor change.	The NRVA methodology is improving with each iteration as is its analytical product. Initially, given its nationwide focus, there were sampling errors and inconsistent collection on some indicators. It is very transparent and can easily be disaggregated. But its focus on districts and villages makes it more vulnerable to measurement errors.
Asia Foundation Survey	The Asia Foundation survey is organized by sector, which is of limited use in systems-based approaches. However, the raw, longitudinal data can be informative.	The Asia Foundation survey is done annually and has been completed X times.	The second iteration of the Asia Foundation survey saw a decrease in possible measurement errors (primarily sampling size, survey size and selection bias) but significant issues remain, especially the more the data is disaggregated at an increasingly granular level.
National Solidarity Program (NSP)	Given its national focus, M&E data of from the NSP could be a useful to develop comparison groups. While not stabilization specific, NSP utilizes survey data in its impact analysis that could inform other approaches.	NSP reporting data is comprehensive, continuous and the impact evaluation data is generated consistently. There is good baseline data and the reporting and impact data collection is reliable.	NSP has not yet expanded into more unstable areas, and selection biases are problematic.

DATA SOURCE	APPLICABILITY	FREQUENCY	RELIABILITY
SIGACTs	SIGACT reporting will provide a useful data point for tracking incidents of violence against expeditionary forces, though it must be used carefully, as higher or lower rates of SIGACTs likely omit important stability factors.	SIGACTS are updated continuously on CLASSIFIED systems. However, the unclassified SIGACTS dataset ends in December 2010.	SIGACTs reporting is consistent and reliable where expeditionary organizations are present in sufficient numbers. But SIGACTs correlate to the presence of Western troops and civilians. In and of themselves they say little about overall stability, and must be carefully combined with other indicators.
USAID M&E Efforts	This data set covers USAID M&E efforts outside Afghan Info and includes third party evaluations. These data sets are outcome/impact oriented but sometimes do not meet appropriate standards of rigor.	The USAID Evaluation Policy helps lay out criteria for when outcome oriented/third party evaluations are required, but these evaluations are sporadic and usually singular reports rather than ongoing efforts.	Most of these evaluations suffer from a lack of baseline data and are often dependent on few data points from surveys, focus groups, and interviews.
Afghan Info	Afghan Info could be helpful in tracking implementing partner activities in specific geographic areas. Afghan Info contains performance related data. Its use in determining impact is limited.	Afghan Info is continuously updated	The performance data in Afghan Info can be disaggregated by technical sector, type of activity, and geographic area down to the district level. It contains GIS data as well. Data added prior to 2009 is of poorer quality than newer additions.
Commander's Emergency Response Program Data (CERP)	Given the size and scope of CERP, and its predominant security focus (as opposed to development) its data can help track externally financed efforts that provide comparative analysis to USAID funded stabilization projects in the same area.	CERP data is primarily performance oriented, and is updated on a continual basis.	CERP budgetary reporting is questionable. Its performance reporting is subjective and unreliable. Actual impact data for CERP is virtually non-existent.

DATA SOURCE	APPLICABILITY	FREQUENCY	RELIABILITY
Afghan National Quarterly Assessment Research (ANQAR) Survey	National poll of ca.13000 Afghan households focused mostly on security issues but also covering some governance and economic issues.	Quarterly	The ANQAR survey is long. Some questions have a high frequency of "I don't know" responses indicating survey fatigue. There are significant reliability challenges with this data source.
BINAA	A survey of ca. 13000 Afghan households across ca.80 Key Terrain Districts (KTDs) conducted by the IJC. Survey focuses on security and governance issues.	Quarterly	The BINAA survey is long and subject to invalid aggregation techniques and inadequate sampling sizes. It has a high frequency of "I don't know" responses, suggesting respondents may be suffering from survey fatigue. There are significant reliability challenges with this data source.

FINDINGS FROM EVALUATIONS OF STABILIZATION INTERVENTIONS

Examining the previous stabilization evaluations and what they tell us about these methods, frameworks, and data's applicability and effectiveness in Afghanistan will inform what monitoring and evaluation of stabilization efforts should look like in the future.

Afghanistan Stabilization Initiatives (ASI)

The Afghanistan Stabilization Initiative's objective was to address instability by fostering and strengthening conditions that build links between the Government of the Islamic Republic of Afghanistan (GIRoA) and local Afghan communities. ASI's M&E program consisted of three levels of analysis and data: The first level consisted of evaluating specific initiatives and their outputs. The second level analyzed the impact of those individual activities, particularly on stabilization. The third level of analysis encompassed evaluating overall stability in the district with the intention of trying to demonstrate plausible contribution from individual activities to the level of stability in a district.

ASI undertook multiple lines of effort to monitor and evaluate individual activities at the first two levels of output and impact. ASI's staff members monitored and evaluated activities through field observation. ASI-East staff used the DSF, and ASI-South staff used the RSSA (and parts of the DSF) to further support both programming and M&E. ASI also hired the company Altai to conduct third party evaluation of activities. Using focus groups, perception surveys, mini-surveys, site visits, case studies, and key informant interviews, Altai provided reports on the impact of individual activities on a district's stabilization. A key finding of these efforts was that ASI-East's use of DSF to program activities against sources of instability had positive results because it provided a coherent method for allocating resources, and justifying this allocation, to particular activities and areas according to stability objectives. However, identified sources of instability were frequently not specific enough to provide sub-district-level objectives and indicators.

The Altai study also noted that activities should be evaluated in clusters with their impacts monitored through case studies combining observations and quantitative and qualitative research. The Altai study can be considered "systematic" according to the above methodology discussion as it lacked a baseline and had no methodology for selecting treatment and control locations or causal attribution. The intensive nature of this multi-method evaluation effort limited the speed in which data could be evaluated or desired data could be collected.

To monitor and evaluate the third level, ASI used two different tools for evaluating the overall stability in a district. ASI used Altai to conduct surveys that would be used to determine the level of stability. At the same time, ASI-East hired RSI to collect overall stability indicators using individual interviews, surveys and direct observation. The seven indicators were: 1. Recognition of the district government; 2. Civilian security; 3. Market activity; 4. ANSF presence; 5. Freedom of movement; 6. Perceptions of governance; and, 7. Perceptions of security.

RSI used multiple methods for collecting and verifying these methods. Below is a paraphrasing of RSI's methodology as they described in their report to USAID:

- **A large-scale quantitative survey** using a multi-stage cluster probability sampling approach in 10 districts. The primary sampling units (PSU) were clusters of villages 5km, 10km, 15km and 20km from t e district center, and the secondary sampling units (SSU) were randomly selected villages within those clusters. Within villages, households were chosen through "random walk method" and in very insecure areas, randomly with the help and under the protection of village Maliks and Mullahs. At the household level, stratification by age group was achieved in the following categories: 15-24, 25-40, 41-59 in order to achieve greater precision. A total of 20 households were interviewed in each village "cluster". About 23% of all known village "clusters" were chosen in each district. In estimating sample sizes necessary to obtain a representative sample in each district, RSI employed a standard proportional approach (between .5% and 1.5% per district),5 with larger percentages for districts with fewer people in order to raise the overall confidence levels of the survey results for that district. The high proportion of clusters chosen (23%), coupled with the assumption that different clusters in particular geographic areas are homogenous, gives a high degree of precision to the results. These populations resulted in a 95% confidence level with a margin of error (ME) of between .035 and .05 (3.5% and 5%) per district.
- **Twenty-five Focus groups**. Focus groups included separate groups for women and others segregated between younger and older men.
- **District Government Center survey.** Enumerators sat inside or just outside the district center for one month, and asked people entering the district center office, why they were there and other related questions. This was carried out in nine districts (Barmel office was closed).
- **Market center survey.** Enumerators first made a list of all shops in the market stratified by type, and then interviewed a few of each shop type, according to a convenience sample. Shoppers in the market were also observed and the overall market atmosphere recorded.
- **A government officials' survey**. Government officials in nine districts were targeted for the survey. These included district government staff, ANA Commanders and staff, teachers and in some cases, senior local elders and Maliks. All district government officials were interviewed within their own districts.

- **A driver survey**. A convenience sample of commercial and private drivers in bus stations.
- **Data validity and verification processes:** A number of data validity and verification processes were followed including GPS grids for the majority of villages, photographs where GPS grids were not available, respondent phone numbers, field corroboration, data triangulation, and quality control on the survey forms. There was clear anxiety about answering some of the questions in the surveys, resulting in a high non-response rate for the large-scale quantitative survey, particularly on matters relating to the Taliban and sometimes to ISAF.

This RSI evaluation provided an overall stability picture at the district level that provided situational awareness for programming. However the study was not designed to test whether or not ASI program activities had stabilizing effects. The RSI study could serve as a baseline for future programming in the surveyed districts. One key point that it highlighted as that many of the stability issues are linked directly to rule of law or transnational issues that are unlikely to be influenced by any one program or project. This supports the need to have a comprehensive evaluation effort incorporating all of the components of stabilization initiatives, since the result of one program may depend on other initiatives.

Quasi-Experimental Evaluations of State Building in Colombia

USAID/Colombia attempted to design and implement an experimental design to evaluate the impact of the Mission's state-building and consolidation program. The Mission wanted to understand the effects of military and civilian institution expansion into new territories on a variety of dimensions (on order and stability, development, economic integration, and on democracy).[29] After an arduous three-year process, the evaluation team finally settled on a hybrid design that employed a mix of survey and observational data collection that tracks household and community-level impacts. This quasi-experimental design uses representative samples of control and treatment municipalities through block matching techniques to control for variables between municipalities and also incorporates staggered baselines.[30] Key stakeholders involved in this evaluation have privately commented on the constraints they faced during this process. In particular, they observed that USAID's institutional incentives for promotion and career advancement may have prevented the taking of a technical risk to support a novel, yet rigorous experimental design in this evaluation.

National Solidarity Program (NSP)

The National Solidarity Program (NSP) provides an example of a multi-year field-based randomized control trial performed in Afghanistan. In 2007, the Ministry of Rural Development (MRRD) and World Bank commissioned a team of researchers from Harvard and MIT to undertake an independent impact evaluation of Phase II of the NSP (NSP-II). The study compares 250 treatment communities with 250 control communities in ten districts in Balkh, Baghlan, Daykundi, Ghor, Herat, and Nangahar provinces.[31] It was designed to report impacts at different stages of assistance project cycles. The ten districts were selected based on size, security conditions, and none of the districts could have previously participated in NSP.[32] The study conducted to quantify and explore changes across indicators including economic

[29] Steele, Abbey. An Evaluation of Statebuilding in Colombia. December 9, 2011

[30] Ibid

[31] Adreskan, Gulran, Farsi, Chest-e-sharif, Balkh, Khost-wa-freng, Sang-e-takht, Dulina, Hesarak and Sherzad districts

[32] Andrew Beath, Fotini Christia and Ruben Enikolopov, Randomized Impact Evaluation of Afghanistan's National Solidarity Program – Final Report, p. 2, February 14, 2012

activity, agricultural production, access to infrastructure and services, and structures and perceptions of local governance.[33] The team used 199 constituent indicators to test fifty hypotheses.

The baseline survey was conducted in August and September 2007 and involved nearly 13,000 respondents. A matched-pair cluster randomization procedure was applied to pair treatment and control villages. In order to minimize the probability of spillovers biasing estimated impacts, villages within close proximity to each other were clustered and assigned a "treatment" or "control" status, though the large number of clustered villages in some districts precluded the assignment of the same treatment status in 17 village clusters.

Following the baseline survey, midline and endline surveys were conducted respectively between June and October 2009, and May and October 2011. Four survey instruments were used to interview male heads of household, female villagers, and male and female village leaders. The interviews were conducted in a variety of forums including household interviews and focus groups. The attrition of villages due to security conditions and other reasons meant that some village pairs had to be excluded from the analysis in order to preserve the internal validity of the experiment.[34]

In the final analysis, the NSP evaluation showed that the program had positive impacts on perceptions of the Afghan government, security, and ability to meet basic needs. There was also some evidence that NSP reduced the frequency of violent incidents in the vicinity of project villages. However these impacts were not seen in areas where violence was already frequent prior to the implementation of NSP. The NSP program should thus be considered and effective in areas where stability is sufficient for development programming. However, NSP was not designed to be an effective stabilization program and had no impact in relatively unstable areas.

Local Governance and Community Development Program (LGCD)

This DAI-run program had two primary objectives, increase GIROA's ability to provide services and increase local communities ability to meet basic needs. "Community ownership" and identifying "root causes" of instability were implementation watchwords. Over USD 110M dollars was disbursed in total across all regions (22 of 34 provinces). Initiatives included close work with various U.S.-led Provincial Reconstruction Teams (PRTs) in volatile RC-E. Activities ranged from health, to infrastructure, to civil service capacity building. As both a "top down" and "bottom up" program, the impact on local populations was measured by random sampling of villages. No attempt was made to identify "direct" beneficiaries; enumerator selected survey respondents randomly in project villages and non-project villages to measure whether respondents perceived that the local environment had become more or less stable. In FY 2011, 42 percent of respondents reported stability in their respective communities had improved, which exceeded the target of 35 percent.

The LGCD survey was conducted in four iterations over 2010-2011, with a total of 5,411 respondents selected randomly for interviews in 123 different randomly selected villages, across 64 districts, in 20 provinces. In total, 110 villages witnessed an LGCD activity that was either underway or completed in the three months prior to survey fieldwork. These villages formed the "treatment group" in contrast with 13 villages that formed the "control group" where no activity was conducted.

[33] NSP Impact Evaluation – Summary and Update, Andrew Beath, April 28, 2009

[34] Andrew Beath, Fotini Christia and Ruben Enikolopov, Randomized Impact Evaluation of Afghanistan's National Solidarity Program – Final Report, p. 11, February 14, 2012

LGCD used multilevel regression models of the survey data, plus M&E indicators and other project activity data, to understand which project activities made respondents more likely to report improved stability. Multilevel modeling works by partitioning the variance in the stability variable between individual, village, and district levels of clustering. That is, the individual survey respondents are grouped into villages, and villages are grouped into districts, and the model estimates the degree of correlation between individual perceptions of stability within each village and district. The degree of correlation between responses shows the effect of local context on perceptions of stability.

Individuals in villages were 17 percent more likely to give the same answer to the stability question than the average across all respondents. Similarly, individuals in the same district were 11 percent more likely to give the same answer to the stability question. LGCD hypothesized that its project activities were the key factor that improved stability locally. The hypothesis was validated; adding project activities to the model reduced the 17 percent "village effect" to zero. LGCD activities explained why people in project villages were more likely than average to report improve stability.

Disbursements of program funds, both the total disbursement and the rate of spending, were shown to increase perceived stability. Disbursements on underway activities, relatively small Local Stability Initiatives (LSI) designed to mitigate local conflict, and activities with community contribution increased stability by 5-8 percent depending on the project. Activities within a 1km radius of a village had a larger impact than activities within a 2-5km radius. This finding helped sort the LGCD impacts from the potential impacts of other nearby projects. The findings showed diminishing returns to stability with larger programmatic disbursements. Overall, the findings suggest that biggest impact on stability arises from LSI activities with community contributions that disburse more funds over shorter periods of time.

The perception that GIRoA's ability to provide security improved was the most important predictor of improved stability. Improved government services and government responsiveness in general were also strong predictors of improved stability. More specifically, government provision of potable water, road improvements, and improved irrigation were also significant predictors of stability. Irrigation projects were particularly effective when implemented with community contributions.

Community Development Program (CDP)

The CDP-N program was a 12-month program designed to assist urban and rural households in nine northern Afghanistan provinces. Increasing local resources to meet basic needs and improve economic and community resiliency were primary goals. Stabilization and better relations between communities and government were later identified as additional goals.

Evaluation of program effectiveness included specified geography (five projects in each province). Provincial and district-level was conducted, as well interviews with key elders and some direct beneficiaries. Notably, at least one women's project per district was also reviewed for impact.

Cash for work efforts were found to be limited in impact and with only short-term impact. Some tension was created between beneficiaries and non-beneficiaries. Elders were most complimentary about projects focused on community assets, though some criticized the lack of an enduring impact. One unintended outcome centered on local communities linking benefits with the local NGO or implementing partner – and not from the government. Note: the only role for government was project approval, which complicated any linkage between local officials and the targeted community.

Helmand Monitoring and Evaluation Program (HMEP)

This Helmand PRT-initiated program focused on Helmand-wide stabilization efforts, in one of Afghanistan's most violent regions, where a large UK force contingent served alongside even more U.S. Marines. The program assessed effectiveness across eleven districts. HMEP targeted information to determine the readiness of the province to move into "transition" phase, in line with GIORA and ISAF priorities. (Note: Helmand capital, Lashkar Gah, fell into an early, high-visibility "hand-off" category to ANSF in the security sector.)

HMEP focused data collection in three areas measured against an overarching Helmand Plan: transition readiness, people's priorities, and what works best. Findings were then plotted against measurements that included, security, support for the Taliban, governance, corruption and development. HMEP data showed disconnects between Helmand residents and their government, perceived to be non-accountable in that they are non-elected. Notably, more contact with government did not translate into a more positive correlation. Data also showed persistent criticisms of governmental lack of capacity and corruption. On counter-narcotic efforts – a key indicator given Helmand's central role in poppy cultivation – data revealed skepticism toward governor-led eradication programs. In the health sector, respondents prioritized improved quality over more access. In education, emphasis placed on improved training for teachers and resources. The local population associated functioning, secure bazaars as biggest indicator of government success. The same held true for infrastructure improvements, especially electricity, and agricultural assistance (though corruption was cited in the wheat seed distribution program). Rule of law concerns likewise centered on questionable training and perceived corruption. GIROA outreach via radio messaging appeared most effective, alongside close communication with religious and tribal leaders. Overall, the local population cited a decrease in sense of security but an increase in the local economy, which likely reflected the sudden increase of U.S. resources (CERP) in areas that U.S. Marines cleared beginning in mid-2009.

The reliability of HMEP's findings stems from its diverse information sources, which included: PRT data, RC(SW) (U.S. Marine-led command) input, and the HMEP head of household survey. Data was also collected via "secondary sources" and "other surveys." One notable strength in data collection was the equitable percentage of female respondents (about half of each sample pool). An inherent challenge in Helmand remains persistent influence of narco-trafficking and the opium economy. While HMEP could point to successes, the governor's effort on behalf of GIRoA to eradicate poppies likely undercut other initiatives. HMEP did not measure these associated effects in Afghanistan's top poppy producing region. Data might also have been measured against disproportionate levels of U.S. versus UK resources (CERP, for example, among deep-pocketed Marine commanders). The influx of troops following the U.S. decision to surge troops into southern Afghanistan did not appear to convince most residents of an enduring improvement in security gains. That indicator, not flagged, would not bode well for any security transition (Afghan-lead) timeline.

THE WAY FORWARD: DESIGN PRINCIPLES FOR STABILIZATION M&E BASED ON LESSONS LEARNED

Establish Baseline Data and Adequate Data Sources

Since stabilization programs seek to prompt a change in behaviors (making individuals and communities stakeholders in stability and shifting the behavior of destabilizing actors), and perceptions drive human behavior, it is important to measure population perceptions and behaviors as early and directly as possible. While reliable quantitative data is difficult to gather in high risk environments like Afghanistan, it is possible. Randomized control trials are the most scientific means of attributing causal relationships between projects, programs, policies and outcomes of interest.

Add Remote Observations and Integrate Them into the Analysis

While direct measurement of local perceptions and behaviors is critical, perception data is all the more rich when it can be integrated with, and analyzed alongside, remotely observable indicators of population behavior. These include such things as movement patterns, price fluctuations, cellphone usage patterns or participation rates in specific programs. These so-called "honest signals" provide a context for evaluation of directly collected perception data. For example, social scientists at the Defence Advanced Research Projects Agency work closely with computer scientists to gain an environmental understanding of stabilization in Afghanistan by reviewing large data sets. Their ability to ingest, format, manipulate, and extract patterns and signatures from vast datasets creates new opportunities for impact evaluation and monitoring in the future. These tools also enable non-intrusive assessment in places that are simply too risky for traditional on-the-spot evaluation. Analysis of remotely observed signatures can inform judgments on the stability of a community over time.

Use Flexible, Mixed-Method Approaches

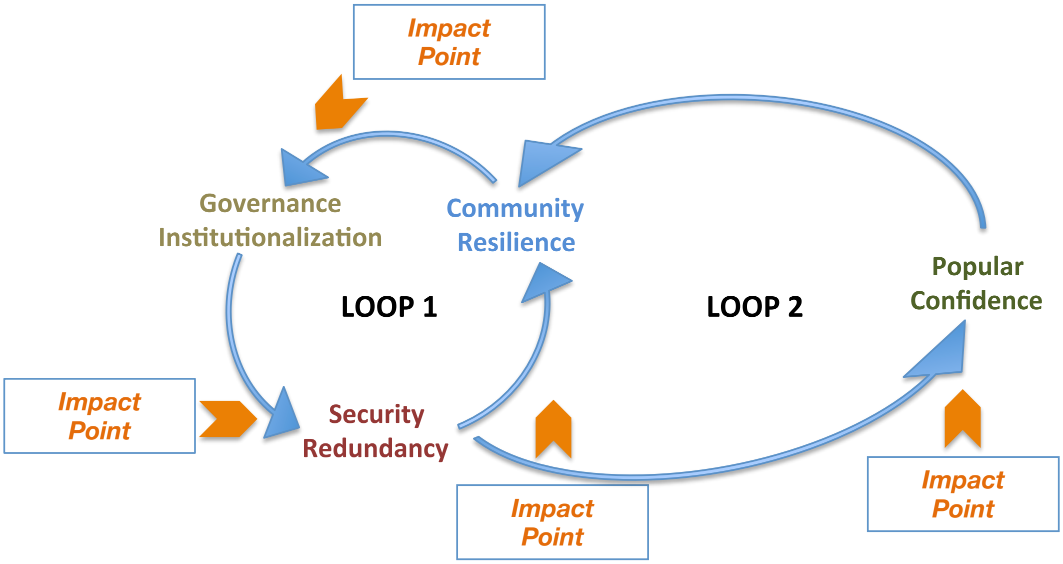
While the use of advanced quantitative design and remote observation is ideal for conducting impact evaluations, these methods may not always provide useful insights for implementers. Mixed-methods approaches that recognize the limits of these data sources in conflict zones and take into account the analytical constraints of these environments can add much power to understanding the environment under study.

In order for donors to develop a meaningful understanding of the emergent socio-political dynamics among communities in Afghanistan, they must develop new capabilities for measuring Afghans' perspectives of security, conflict dynamics and local disputes, and support for government initiatives and institutions. Sensing tools that combine surveys with remote observation with qualitative field research can facilitate deep contextualized understanding.

Use Theories of Change and Link Them to an Accurate Diagnosis of How Things Work in the Environment

Just as drivers of instability affect entities within the system, projects must seek to affect system dynamics, and the causal linkages among them, through an explicitly articulated theory of change. These "if-then" statements not only clearly articulate the expected impacts of interventions but by explicitly connecting theories of change to impact points within the system, planners and implementers can develop a much more sophisticated understanding of how projects' impacts may produce follow-on effects, and accelerate improvements in stability.

FIGURE 5: DOUBLE-LOOP STABILITY MODEL WITH IMPACT POINTS



This simplified model depicts the overall systems logic–the "way things work" in a given system–to allow implementers to build detailed impact models for specific projects. The impact points represented in this figure correspond with specific project interventions (with defined theories of change that are aligned to key nodes of the system) and corresponding measurement approaches. Such diagnoses are necessarily (and deliberately) simplified descriptions of intensely complex and nuanced real-world dynamic systems. There have been attempts to comprehensively capture complexity (see figure 3), but it is easy to see how such comprehensiveness leads to the same problem as failing to diagnose the system at all—it is impossible to tell what is important and what is not.

Use Theories of Change and Link Them to an Accurate Diagnosis of How Things Work in the Environment

The difficulties of evaluating stabilization programs are challenging, but there are ways of mitigating the most harmful effects of those challenges. Smart, streamlined analysis based on understanding the systems and how issues relate to each other can help to identify nodes that need to be influenced. Sometimes these nodes are sources of instability and resiliencies. The desired effects of interventions should be outlined in theories of change. Even anecdotal methodologies can become more rigorous when linked directly to analysis of how they impact or do not impact theories of change.